PhytoKeys 267: 345–360 (2025)
DOI: 10.3897/phytokeys.267.172490

Research Article

Determining the best morphological characters for taxonomic identification of *Inga* species in the Colombian Andes using correlation, entropy and a discriminant index

Kelly T. Bocanegra-González[1,2], Thomas S. Prieto-González[1], Martin Pullan[2], Kyle G. Dexter[2,3,4], Hannah Atkins[2], R. Toby Pennington[2,5,6]

1 *Research Group on Biodiversity and Dynamics of Tropical Ecosystems, University of Tolima, Ibagué, Colombia*
2 *Royal Botanic Garden Edinburgh, Edinburgh, UK*
3 *School of GeoSciences, University of Edinburgh, Edinburgh, UK*
4 *Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy*
5 *Department of Geography, University of Exeter, Exeter, UK*
6 *Department of Biology, Washington University in St Louis, St Louis, USA*

Corresponding author: Kelly T. Bocanegra-González (ktbocanegrag@gmail.com)

OPEN ACCESS

Academic editor: Patrick Herendeen
Received: 18 September 2025
Accepted: 14 November 2025
Published: 12 December 2025

Citation: Bocanegra-González KT, Prieto-González TS, Pullan M, Dexter KG, Atkins H, Pennington RT (2025) Determining the best morphological characters for taxonomic identification of *Inga* species in the Colombian Andes using correlation, entropy and a discriminant index. PhytoKeys 267: 345–360. https://doi.org/10.3897/phytokeys.267.172490

Copyright: © Kelly T. Bocanegra-González et al. This is an open access article distributed under terms of the Creative Commons Attribution License (Attribution 4.0 International – CC BY 4.0).

Abstract

The genus *Inga* (Leguminosae, Caesalpinioideae) is a key component of tropical American forests, yet its high species diversity and morphological complexity hinder accurate identification, especially in montane environments. In this study, we evaluated 49 morphological characters (32 categorical and 17 continuous) from 73 *Inga* species occurring above 1000 m in Colombia. We applied a multi-criteria framework combining Spearman correlation, Shannon entropy and a discriminant index to prioritise diagnostic characters, based on redundancy, variability and discriminatory power. The results indicate that categorical characters, particularly stipule shape, indumentum type on both lamina surfaces and the petiole, interfoliar nectaries shape, leaflet apex and base shapes are the most informative for species identification and therefore received the highest scores. In contrast, continuous characters showed limited variation among species and were less useful and received lower scores. A usability test demonstrated higher accuracy and faster identifications when using the highest-scoring characters. This framework not only improves identification efficiency in field and herbarium settings but also offers a scalable model for intelligent, user-friendly taxonomic tools. As a complementary output, we present an open-access, interactive identification key, designed to improve the quality and reliability of *Inga* identifications across the Colombian Andes.

Key words: Character prioritisation, Colombian Andes, *Inga*, interactive key, taxonomic informatics

Introduction

The genus *Inga* Mill. (Leguminosae, Caesalpinioideae) is a key component of tropical American forests, characterised by its high diversity (ca. 280 accepted species) and ecological significance (Pennington 1997; Endara and Jaramillo 2011; Palow et al. 2012; Bocanegra González et al. 2025). It is widespread

across a range of tropical moist ecosystems. Most species occur below 1800 m above sea level, although some have been recorded at elevations exceeding 3000 m. These high-elevation species are particularly poorly explored, with limited information on their identity and distribution.

In recent years, important progress has been made in Colombia studying the genus (Romero and Alba-López 2005; Bocanegra-González et al. 2024, 2025). However, research on *Inga* faces several challenges, particularly in the high Andes (over 1000 m). These challenges arise from the poor knowledge of morphological variation amongst species and the absence of specific diagnostic characters that simplify identification and standardise identification keys.

According to Bisby (1970), a character is a concept shaped by criteria such as: *i*) morphological relevance, which ensures a link to plant structure; *ii*) variability, which highlights significant differences amongst species; *iii*) stability over ontogeny and under varying environmental conditions; *iv*) functionality, which reflects the plant's adaptation and survival through the character; *v*) homology over homoplasy, which prioritises characters related to evolutionary relationships; *vi*) simplicity, which favours fewer, but more relevant characters in identification; and *vii*) accessibility, which ensures characters are observable and measurable.

For identifying *Inga*, most authors have used sets or subsets of ~ 74 morphological characters, both continuous and categorical (Leon 1966; Pennington 1997; Romero 2005; Romero and Alba-López 2005; Dexter et al. 2010; Genero et al. 2017; Soto et al. 2021; Bocanegra-González et al. 2024, 2025). These include the number of leaflet pairs, size of leaflets, presence and type of indumentum on both adaxial and abaxial leaf surfaces, wing presence and shape, nectary shape and diameter, inflorescence type, position and fruit type, amongst others (Suppl. material 1). Authors have often combined these characters differently (Leon 1966; Pennington 1997; Romero 2005; Romero and Alba-López 2005; Dexter et al. 2010; Genero et al. 2017; Soto et al. 2021; Bocanegra-González et al. 2024, 2025) and there is no general consensus on which characters to use for identifying species, at least in most cases. It is unlikely that a non-specialist would be able to manage the entire set of characters. In addition, some of these features are plastic and show high levels of variation within and amongst *Inga* species, geographical range, making field and herbarium identification challenging (Bocanegra González et al. 2025), contributing to a high percentage of misidentification in the genus (Dexter et al. 2010; Baker et al. 2017). Moreover, as for many tropical hyperdiverse groups, there is a lack of identification tools and traditional dichotomous keys are often not readily available to users (Särkinen et al. 2025).

Determining the most useful characters for identification can improve the efficiency and accuracy of identification in the field, in herbaria and when constructing identification keys (Bisby 1970; Talent et al. 2014; Rakouth et al. 2025). It can also enhance the performance of numerical procedures, such as artificial intelligence, for species identification (Mäder et al. 2021; Hart et al. 2023).

To reduce the dimensionality of highly multivariate, morphological data and improve numerical identification procedures, characters are often ranked, weighted or grouped (Talent et al. 2014). Importantly, the efficiency gained by prioritising characters may vary depending on the type of identification key being used. In dichotomous keys, where the path is fixed by the author, careful selection and ordering of characters is critical, as early mistakes can mislead the user

in the identification process (Hagedorn et al. 2010). In contrast, multi-access keys (also known as polyclaves) allow users to choose the characters and may tolerate user scoring errors better and also allow identification when certain characters are unavailable (i.e. flowers and fruits) (Ohkawa 2000). Thus, character prioritisation has distinct implications depending on the key structure.

Approaches for character selection and species identification consider characters either collectively or in pairs, facilitating the construction of such keys (Talent et al. 2014; Rakouth et al. 2025). Some numerical methods involve frequency of divisions, entropy reduction, balanced characters, conflict reduction, error minimisation, time efficiency and weighted combinations. These approaches are based on set theory, Shannon entropy, proportional optimisation, discrimination analysis, probability, decision theory and multi-criteria optimisation (Lubischew 1962; Eades 1965; Hall 1967; Bisby 1970; Izenman 2008; Talent et al. 2014; Łysko et al. 2022). Despite the wide range of available methods, none has been universally or consistently adopted as a standardised approach. In addition, methods are rarely used together or compared.

Some of the commonly used numerical methods for character prioritisation reported in literature are Pearson's correlations, aiming to reduce redundancy amongst character sets, that is, the overlap in characters that can complicate the identification process. However, it is not the best approach in numerical taxonomy and ecology, primarily because it measures linear relationships, while biological variation is often non-linear and multidimensional (Eades 1965; Hall 1967). Additionally, Pearson correlation analysis is sensitive to outliers, assumes normality in the data and loses information in categorical characters (Schober et al. 2018). Another correlation used is the Spearman's, which measures the strength and direction of a monotonic relationship between two variables without assuming a specific frequency distribution. Unlike Pearson's correlation, it does not require a linear relationship or interval-scale measurements and works for ordinal data (Hauke and Kossowski 2011).

Shannon entropy has been common in character selection (Eades 1965; Hall 1967; Bisby 1970; Talent et al. 2014) and can be understood as measuring the variability of elements within a given set, quantifying the number of yes/no questions needed to identify an element in a set (Carcassi et al. 2021). This method preserves more information about the distribution of data by avoiding the simplification that results from averaging values. Shannon entropy can be applied to non-linear systems, but it provides only a partial view of their complexity. For this reason, it is often recommended to use entropy in combination with other statistics, which can capture additional aspects of non-linear biological dynamics (Eskov et al. 2017).

Another common approach is linear discriminant analysis (LDA), which was developed to solve taxonomic (i.e. classification) issues using multiple botanical measurements (Brown and Wicker 2000; Izenman 2008). LDA assesses which characters identify unique species and has been used in multiple studies and is presently a popular method for training machine-learning (Lubischew 1962; Łysko et al. 2022). However, LDA assumes normality and requires multiple observations per class to estimate within-class variance, which limits its application when only a single observation per species is available or when data are only available as ranges of a continuous character per species. LDA can be replaced with a univariate discrimination index. This metric evaluates, for each character,

the proportion of species that can be uniquely identified, based on that character alone, thus capturing its stand-alone discriminatory power. While it does not model multivariate separation as LDA does, it is robust for datasets with limited replicates, avoids violations of normality assumptions and still provides a quantitative basis for ranking characters in terms of their identification value.

In this study, we analysed a set of 49 continuous and categorical characters across Colombian Andean *Inga* species to identify those that are most effective for species identification. We employ a combined approach that includes Spearman's correlation, Shannon entropy and a univariate discrimination index.

Our goal is to determine the most relevant set of characters required for identification, thereby enhancing the efficiency and accuracy of *Inga* identification. Finally, we present an interactive key that can be used with all the available characters, either including all characters or focusing on those prioritised as most informative.

Materials and methods

Data collection

Inga species occurrences in Colombia above 1000 m were extracted from the Global Biodiversity Information Facility (GBIF.org 2024) (Fig. 1). These records were used to compile the list of species reported in the region. A total of 1339 records representing 73 species were included; 49 morphological characters for *Inga* (32 categorical and 17 continuous), selected from the 74 characters described in *Inga* studies were used in the analyses (Suppl. material 1). The 49 characters were chosen because they are consistently reported in literature (Leon 1966; Pennington 1997; Romero 2005; Romero and Alba-López 2005; Dexter et al. 2010; Genero et al. 2017; Soto et al. 2021; Bocanegra-González et al. 2024, 2025) and complemented with measurements taken from herbarium collections (COL, E, UDBC, TOLI), where all 73 species were represented and identifications confirmed.

In the case of categorical characters, their definitions were standardised across sources to ensure consistency (Suppl. material 2) and using as a base the plant glossary by Beentje (2010). We also considered the historical usage of terminology for these characters in the genus, especially that established by Pennington (1997). All the information was organised in a spreadsheet for subsequent analyses.

Data analyses

The analyses were conducted in Google Colab, an interactive Python environment. All categorical variables were transformed into numerical values using an average relative frequency encoding method (Butcher and Smith 2020). For each character, the relative frequency of each state was computed by dividing the number of times that state appears by the total number of non-missing entries for that character. The value assigned to a species corresponds to the frequency of the state it exhibits, meaning that species sharing a common state receive the same numerical value. This approach reflects how common or rare each state is within the dataset. For multistate categorical characters, we calculated the relative frequency of each state across the dataset and used their average as the final encoded value. This approach preserved the information content of composite character states, while avoiding artificial inflation of dimensionality.

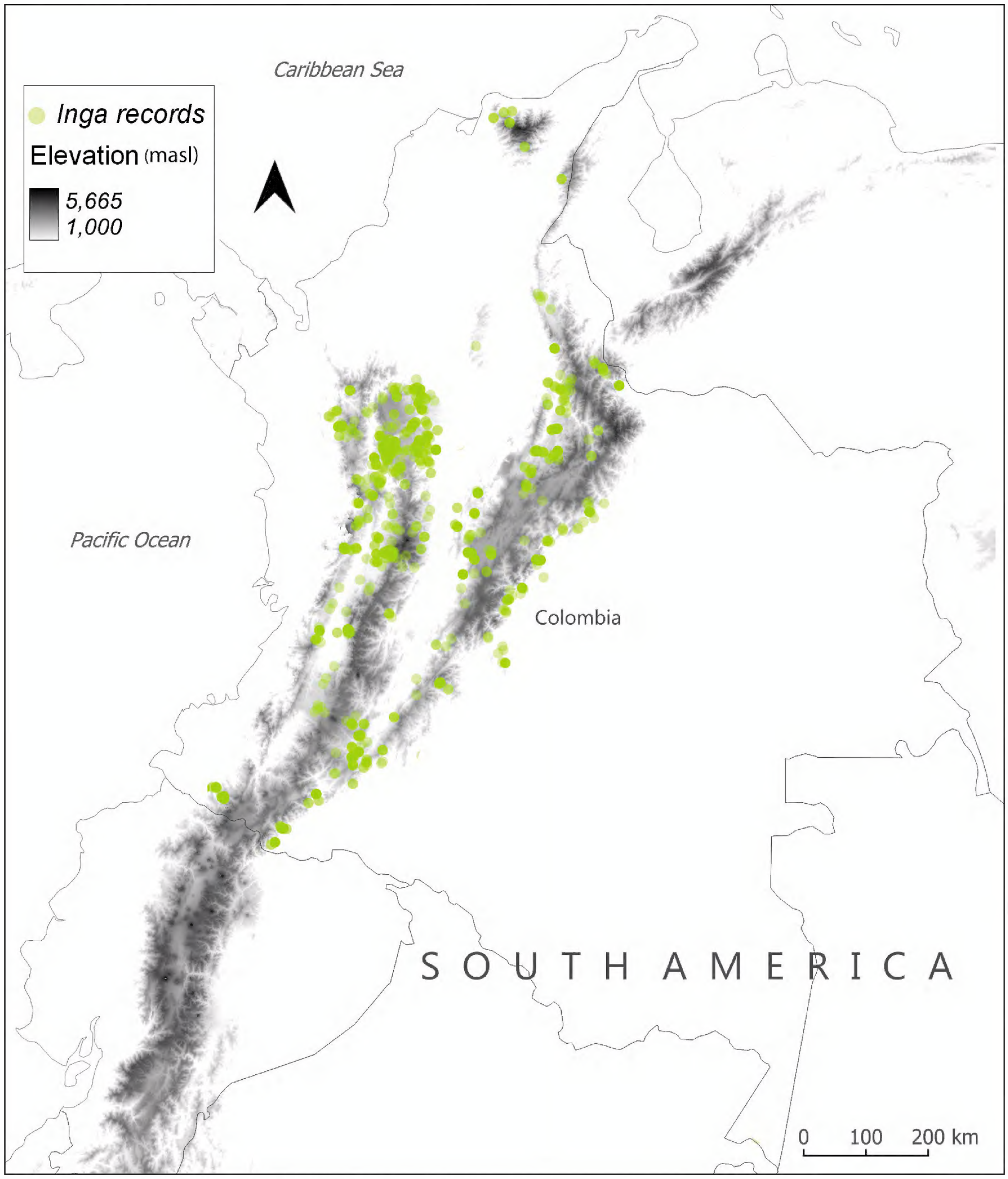

Figure 1. Curated *Inga* collections from the Colombian Andes obtained from GBIF range from elevations above 1000 m to a maximum of 3210 m a.s.l.

In the case of continuous characters, the metrics corresponded to the minimum and maximum values of each state. These were unified to establish ranges and were subsequently discretised into categorical classes using defined intervals, based on their observed ranges and biological relevance. Each continuous value was assigned to a bin. This binning approach facilitated integration with entropy and discrimination metrics. Subsequently, all character columns were normalised to a 0–1 range using Min–Max scaling to ensure comparability.

Pairwise character–character correlations were then computed with Spearman's rank correlation using the "*spearmanr*" function from the "*scipy.stats*" module.

Shannon entropy was calculated for each variable to quantify the distribution of its character states. Higher entropy values indicate greater variability across states and, therefore, more potential for discrimination, whereas lower values reflect conserved characters that separate fewer species. Entropy was computed using the function "*scipy.stats.entropy*", applied to the frequency distribution of states for each character.

Discrimination analysis was conducted using a custom univariate method. For each character, we calculated the proportion of species that could be uniquely identified, based on that character alone. This was done by grouping the dataset by character states and counting how many of those states were restricted to a single species using the "*pandas*" library. The number of exclusive states was then expressed as a percentage of the total number of species, defining the discrimination index as a measure of the standalone discriminative power of each character. A univariate approach was necessary to assess the relative contribution of each character in isolation, since the dataset is structured at the species level, with single composite records per species rather than multiple individuals.

Subsequently, a multiple weighting analysis was performed using three metrics: Spearman correlation, entropy and the discrimination index. These metrics were normalised using the Min-Max scaling method to ensure comparability. Additionally, a bilateral penalty was applied to the correlation metric to discourage both highly redundant and overly uncorrelated characters. Correlations below 0.10 and above 0.60 were penalised linearly, favouring characters with intermediate correlation values that contribute complementary, but non-redundant information to the dataset. Spearman correlation was weighted at 20% to account for redundancy among characters and to prevent highly correlated traits from dominating the analysis. Entropy received the highest weight (60%) because it reflects the overall variability and distribution of states across species, making it the most reliable indicator of consistent taxonomic signal. Finally, the discrimination index was weighted at 20% to incorporate the contribution of characters with uniquely identifying states, while avoiding overemphasis on characters informative only in a few species. Finally, an interactive key was built using Padme (Royal Botanic Garden Edinburgh) with the full character set.

Usability assessment of the interactive key

To assess the practical utility of the interactive multi-access key developed in this study, we conducted a usability experiment comparing species identification performance when using the key with the full character set without prioritisation (unscored list) versus the prioritised characters (scored list). Ten participants with general botanical training, but no prior experience with *Inga* were selected. Each was provided with a set of three specimen photos reflecting observable morphological characters. Participants were randomly assigned to two groups (n = 5 and n = 5). Group A used the unscored list and Group B used the scored one (prioritised order). Each participant identified the same specimens using the same set of 49 characters.

Performance was assessed by comparing the mean identification time between the two-character lists and the number of characters used to reach the identification in each case. Both were evaluated using the Wilcoxon signed-rank test, a non-parametric method suitable for matched samples.

Results

Correlation

The correlations reflect the consistency and biological interdependence of the morphological characters studied, which is logical given the common origin and related functions of the plant structures that showed high correlation (> 60%) (Fig. 2). The consistent relationships between leaflet length and width for terminal

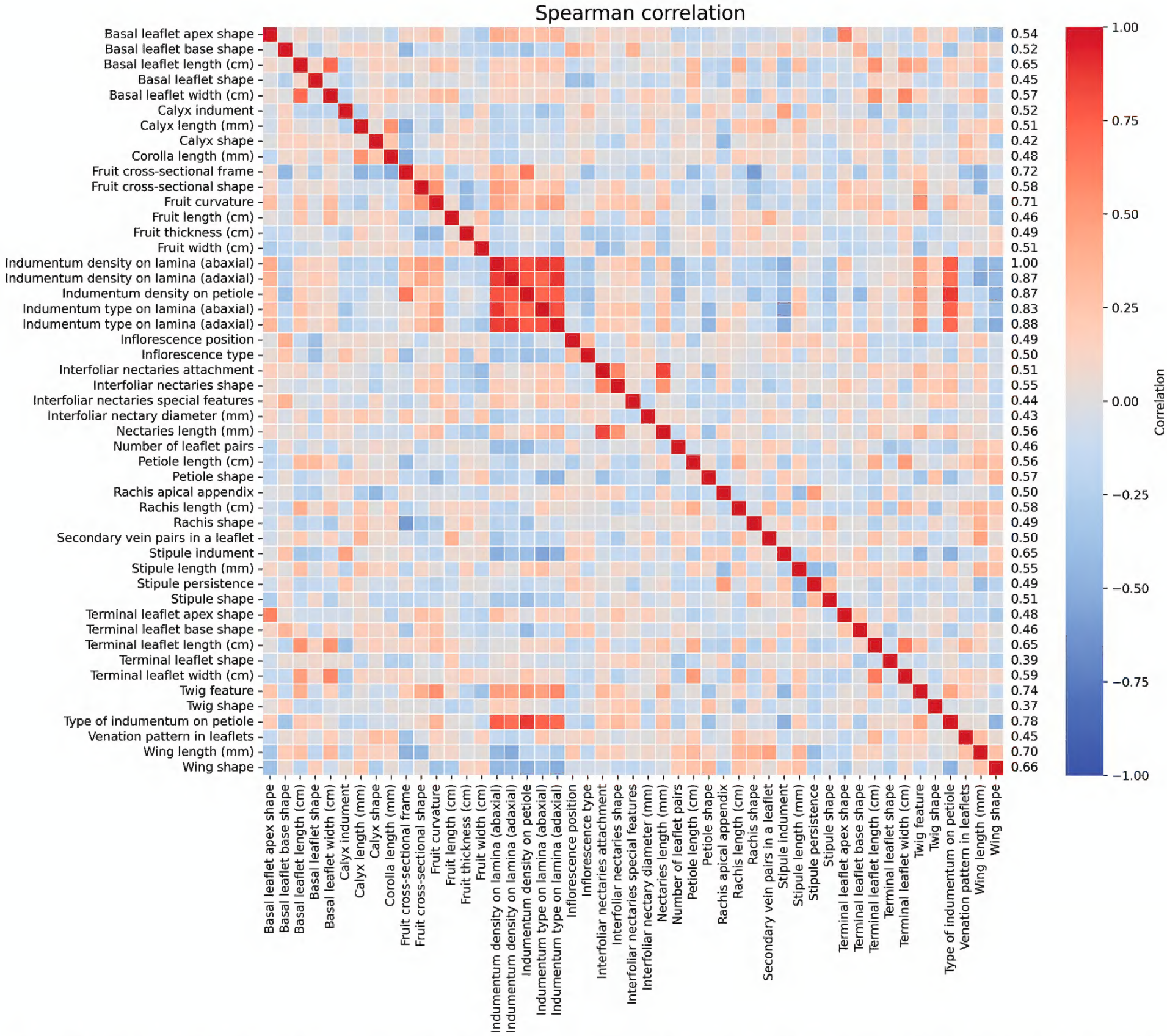

Figure 2. Spearman correlation matrix amongst 49 morphological characters of Colombian Andean *Inga* species. The heatmap shows the strength and direction of correlations between pairs of characters, with red indicating positive correlations and blue indicating negative ones. The values at the end of the axis represent the sum of the absolute correlation coefficients for each character, indicating their overall degree of association with other characters.

and basal leaflets indicate proportional growth patterns. Additionally, correlations between indumentum type and density across abaxial and adaxial lamina surfaces suggest coordinated pubescence (Fig. 2). Similarly, correlation between calyx and corolla length, as well as fruit length, width and thickness, reflect coordinated variation in these characters (Fig. 2). Beyond these groups, most characters showed low to moderate associations, indicating that they vary largely independently and thus provide complementary information for species delimitation in *Inga*.

Shannon entropy

The characters showing the highest entropy values (above 3) are stipule shape, indumentum type on lamina (abaxial), type of indument in petiole, petiole shape and basal leaflet base shape (Fig. 3). This high entropy amongst structural descriptors reflects a wide diversity in how these forms are distributed across *Inga*. In contrast, characters like interfoliar nectary diameter, corolla length, calyx length, inflorescence position and fruit length exhibit lower entropy (below 0.5) (Fig. 3). The lowest entropy is related to characters that are highly conserved across the sampled taxa, which means they do not separate as many species by themselves. It is worth noting that metric characters were transformed into discrete categories in the dataset, which may reduce their range of variation and thus lower their entropy values. Overall, the results reveal a broad spectrum of entropy values from nearly invariant characters to highly variable ones underscoring that only a subset of descriptors contributes strongly to species delimitation. In this dataset, characters with entropy values above 2 clearly form the uppermost group in the distribution and were, therefore, considered the most informative.

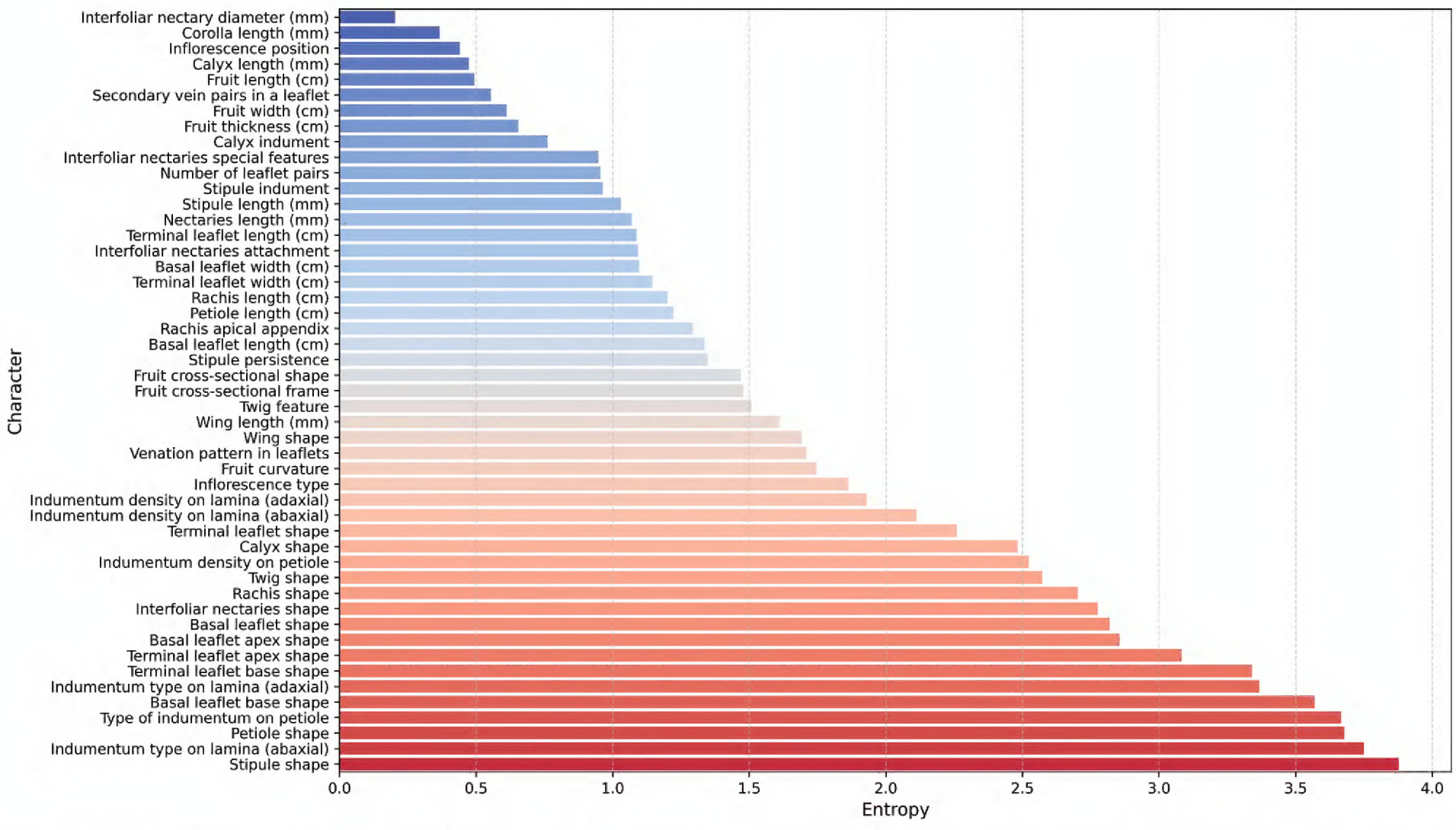

Figure 3. Shannon entropy values for 49 morphological characters of Colombian Andean *Inga* species. The barplot shows the relative information content of each character, with higher entropy values indicating greater variability and discriminatory power across species.

Discrimination analysis

Shape-related characters, such as stipule shape (19.44%), basal leaflet base shape (18.06%), indumentum type on lamina (abaxial) (13.89%), interfoliar nectaries shape and type of indumentum on petiole (12.50%), exhibited the highest discriminant index values (> 12%), highlighting their strong contribution to species delimitation (Table 1). These characters are categorical descriptors that tend to remain stable within species, but differ clearly amongst taxa, making them especially informative for taxonomic identification. A second group of characters showed intermediate contributions (4–11%), including basal leaflet apex shape, terminal leaflet apex shape, terminal leaflet base shape and several additional shape-related features across different structures (Table 1). Although less powerful individually, they still provide relevant discriminatory information. In contrast, most metric characters (e.g. interfoliar nectary diameter (mm), fruit length (cm), corolla length (mm), nectaries length (mm)) consistently scored very low (≈ 1.39% or null) (Table 1), suggesting that continuous measurements, once grouped into size intervals, capture less interspecific variation and, therefore, contribute little to species separation in this dataset. Overall, the discrimination analysis highlights that a limited set of categorical structural characters accounts for most of the univariate discriminative power in *Inga*, supporting species delimitation.

Table 1. Discriminant analysis of 49 morphological characters in Colombian Andean *Inga* species. The table shows the contribution of each character to species discrimination, measured as the Discriminant Index (%). Higher values indicate a stronger influence on the separation amongst species in univariate space.

Character	Discrimination Index %	Character	Discrimination Index %
Stipule shape	19.44	Fruit length (cm)	1.39
Basal leaflet base shape	18.06	Terminal leaflet length (cm)	1.39
Indumentum type on lamina (abaxial)	13.89	Stipule length (mm)	1.39
Interfoliar nectaries shape	12.50	Corolla length (mm)	1.39
Type of indumentum on petiole	12.50	Stipule persistence	0.00
Basal leaflet apex shape	11.11	Interfoliar nectaries attachment	0.00
Terminal leaflet apex shape	9.72	Rachis apical appendix	0.00
Terminal leaflet base shape	8.33	Stipule indument	0.00
Petiole shape	8.33	Nectaries length (mm)	0.00
Rachis shape	6.94	Fruit cross-sectional frame	0.00
Indumentum type on lamina (adaxial)	6.94	Calyx indument	0.00
Inflorescence position	5.56	Indumentum density on lamina (adaxial)	0.00
Inflorescence type	5.56	Indumentum density on lamina (abaxial)	0.00
Terminal leaflet shape	5.56	Fruit cross-sectional shape	0.00
Indumentum density on petiole	5.56	Wing length (mm)	0.00
Twig shape	4.17	Petiole length (cm)	0.00
Basal leaflet shape	2.78	Terminal leaflet width (cm)	0.00
Venation pattern in leaflets	2.78	Number of leaflet pairs	0.00
Fruit thickness (cm)	2.78	Rachis length (cm)	0.00
Fruit curvature	1.39	Basal leaflet length (cm)	0.00
Calyx shape	1.39	Calyx length (mm)	0.00
Interfoliar nectary diameter (mm)	1.39	Secondary vein pairs in a leaflet	0.00
Twig feature	1.39	Basal leaflet width (cm)	0.00
Wing shape	1.39	Fruit width (cm)	0.00
Interfoliar nectaries special features	1.39		

Multiple weighting analysis

The weighted scores indicate that categorical characters concentrate the highest values (Table 2). Stipule shape (1), basal leaflet base shape (0.94), type of indumentum on petiole (0.88) and indumentum type on lamina (abaxial) (0.86) occupy the top ranks, followed closely by petiole shape, terminal leaflet base and apex shape, basal leaflet apex shape and interfoliar nectaries shape (0.74–0.85). These characters emerge as the strongest sources of taxonomic signal once correlation, entropy and discrimination are jointly considered. In contrast, metric characters related to organ dimensions, such as corolla length, leaflet length and width and interfoliar nectary diameter, amongst others, consistently fall to the bottom (< 0.40), confirming their limited contribution under this integrative framework. Between these extremes, there is a group of characters with intermediate scores (0.40–0.73).

Identification characters' guide for *Inga* species, based on weighted scores

The identification of an *Inga* species, according to our analysis, should begin with the highest-value characters, based on the weighted scores (Table 2) and

Table 2. Weighted scores for 49 morphological characters of Colombian Andean *Inga* species, based on a multi-criteria analysis. The score reflects the overall contribution of each character to species identification, integrating four metrics: correlation, entropy and discrimination index. Higher scores indicate characters that are more informative and consistent across species. Characters are ranked from most to least informative.

Character	Weighted Score	Character	Weighted Score
Stipule shape	1.00	Indumentum density on lamina (adaxial)	0.40
Basal leaflet base shape	0.94	Basal leaflet length (cm)	0.39
Type of indumentum on petiole	0.88	Rachis apical appendix	0.38
Indumentum type on lamina (abaxial)	0.86	Rachis length (cm)	0.37
Petiole shape	0.85	Petiole length (cm)	0.37
Terminal leaflet base shape	0.80	Terminal leaflet length (cm)	0.37
Terminal leaflet apex shape	0.75	Terminal leaflet width (cm)	0.36
Basal leaflet apex shape	0.74	Stipule length (mm)	0.36
Interfoliar nectaries shape	0.74	Basal leaflet width (cm)	0.35
Rachis shape	0.72	Interfoliar nectaries attachment	0.35
Indumentum type on lamina (adaxial)	0.68	Nectaries length (mm)	0.35
Basal leaflet shape	0.67	Interfoliar nectaries special features	0.35
Calyx shape	0.58	Number of leaflet pairs	0.34
Terminal leaflet shape	0.54	Stipule indument	0.33
Inflorescence type	0.54	Indumentum density on lamina (abaxial)	0.32
Twig shape	0.53	Fruit thickness (cm)	0.31
Indumentum density on petiole	0.51	Inflorescence position	0.31
Venation pattern in leaflets	0.48	Calyx indument	0.30
Fruit curvature	0.47	Fruit width (cm)	0.28
Wing shape	0.47	Fruit length (cm)	0.27
Wing length (mm)	0.44	Secondary vein pairs in a leaflet	0.26
Twig feature	0.44	Corolla length (mm)	0.25
Fruit cross-sectional frame	0.42	Calyx length (mm)	0.25
Fruit cross-sectional shape	0.41	Interfoliar nectary diameter (mm)	0.20
Stipule persistence	0.40		

progressively move to intermediate and lower-ranked ones. Prioritising characters in this order can efficiently separate the species. Subsequently, less discriminatory characters (those with intermediate weighted scores) should be considered. At the same time, characters with high weighted scores, based on metrics, cannot be directly used unless their biological relevance is interpreted within a specific context. For instance, measuring nectary length (mm) would be irrelevant if the species in question has the more common short cup-shaped nectary. This character would work very well for tubular nectaries.

The multi-access interactive key developed [https://padme.rbge.org.uk/keys/welcome/inga] can be used by selecting any input character; however, the identification process becomes more efficient when using the highest-scored characters. The comparison of mean identification times between the scored and unscored character lists showed no statistically significant difference (Wilcoxon signed-rank test, $p = 0.50$). Nonetheless, participants using the scored list (Group B) tended to complete the identifications more quickly (6.3 min vs. 12.5 min in Group A) and with fewer characters (4.5 vs. 8.3; Wilcoxon signed-rank test, $p = 0.25$). Taken together, these results suggest that prioritising characters can reduce both the effort and time needed for species identification, even if the differences were not statistically significant.

Discussion

This study employed Spearman correlation, Shannon entropy and a discriminant analysis to identify the most effective characters for species identification in the genus *Inga* from the Colombian Andes (above 1000 m), using 49 morphological characters (32 categorical and 17 continuous). This study demonstrates the importance of combining multiple analytical approaches to optimise the selection of diagnostic characters. This is important because no single method is sufficiently accurate on its own in plant identification. This approach reflects a broader trend in modern taxonomy and biodiversity informatics, where no single metric fully captures the complexity of morphological variation. Combining complementary methods, as seen in recent deep learning models for species recognition, improves overall robustness (Carranza-Rojas et al. 2017; Särkinen et al. 2025).

The selected approach highlights the biological consistency (correlation) of certain characters that are closely related in their structural development. It not only optimises identification, but also reduces the time and effort required for data collection in the field and herbaria. For example, in the case of *Inga*, recording the indument type on both lamina surfaces (adaxial and abaxial) is unnecessary, as is recording the separate lengths and widths of the leaflets. While character measurement remains essential for the recognition of *Inga* species, this approach focuses on identifying the most efficient combination of characters to be measured and recorded, avoiding unnecessary redundancy and improving the overall diagnostic process.

Additionally, the Shannon entropy results highlight those categorical characters that are the primary contributors to distinguishing species within the dataset. In contrast, continuous characters show lower entropy and, thus, less discriminative power, likely because their values were grouped into size intervals. Although different interval definitions could be tested, in this study, we

applied a uniform binning scheme to maintain consistency across characters and to align with standard morphological ranges reported in literature.

The discriminant index indicates that a few categorical characters, such as stipule shape, basal leaflet base shape and terminal leaflet apex shape, provide most of the taxonomic resolution in *Inga*. This pattern reinforces that those qualitative structural characters are the most informative for species delimitation in the genus.

Taken together, the previous results provide clear guidance for how characters should be prioritised when constructing identification tools. Furthermore, the choice and order of characters have different implications depending on the type of key. In dichotomous keys, where the sequence is fixed, prioritising stable and easy-to-score characters at the beginning can help prevent misidentifications (Hagedorn et al. 2010). In contrast, multi-access keys allow users to bypass problematic or ambiguous characters, making them more tolerant to variation and observational difficulties (Ohkawa 2000; Särkinen et al. 2025). The methodology proposed here is, therefore, particularly valuable for improving the structure of dichotomous keys, while also enhancing the efficiency of user-driven identification tools. Ultimately, this approach is intended to facilitate accurate and accessible species identification in *Inga*.

The scoring generated here highlights that the vegetative characters are far more important than reproductive ones in the case of *Inga*, reinforcing the need to prioritise vegetative characters where possible. Their relevance is further supported by other studies, where they have proven extremely useful in the development of tropical dynamic keys (Särkinen et al. 2025).

Furthermore, reliance on statistical methods that assume data normality may introduce biases, especially in continuous datasets. Future studies would benefit from integrating these results into the development of automated tools, such as intelligent systems for species identification, based on prioritised characters, which could revolutionise the practical taxonomy of the genus (Wäldchen et al. 2018; Mäder et al. 2021; Hart et al. 2023).

In addition, we highlight the issue of how characters are treated prior to any statistical analysis. The standardisation of characters, no matter how carefully designed, remains partly subjective, as many have undergone changes in their definitions since the most relevant taxonomic treatments of *Inga*, for instance, the type of indu ments (Leon 1966; Pennington 1997). This highlights a persistent challenge: ensuring that, despite conceptual differences across studies, the historical terminology applied to characters can be unified and consistently applied. Such unification is essential to preserve continuity with past work, while enabling comparability and reliability in current and future identification tools. Here, we integrate these characters and refine state definitions, expanding them for species that appeared constrained by characters not well accommodated within the pre-established categories (Leon 1966; Pennington 1997; Romero 2005; Romero and Alba-López 2005; Dexter et al. 2010; Genero et al. 2017; Soto et al. 2021; Bocanegra-González et al. 2024, 2025).

The character prioritisation framework developed also provides a foundation for future integration with AI-based identification tools. Recent advances in computer vision and mobile applications have demonstrated the potential of intelligent systems to facilitate species recognition even for non-specialists (Carranza-Rojas et al. 2017; Wäldchen et al. 2018). Tools such as Flora Incognita (https://floraincognita.com/) already leverage similar principles to guide user input and could benefit from datasets structured around character consistency and discriminatory power.

This positions our results not only as a taxonomic resource, but also as a bridge towards scalable, user-friendly biodiversity tools. This study presents a replicable model for prioritising taxonomic characters in diverse groups such as *Inga*. The interactive key constructed from these results has the potential to facilitate species identification in the field and herbariums, especially for non-specialist users.

Acknowledgements

The authors thank the Herbaria COL, E, TOLI and UDBC. Additionally, we thank all the volunteers who tested the interactive key. KTBG acknowledges the financial support provided by the Sibbald Trust of the Royal Botanic Garden Edinburgh and the Percy Sladen Memorial Fund Grant from the Linnean Society of London.

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Use of AI

AI was used exclusively for language editing.

Funding

This work was supported by Sibbald Trust of the Royal Botanic Garden Edinburgh and the Percy Sladen Memorial Fund Grant from the Linnean Society of London.

Author contributions

KTBG conceived the research and performed the data analyses. KTBG and TSPG collected and curated the taxonomic information. MP developed the interactive key. All authors contributed equally to the writing and revision of the manuscript.

Author ORCIDs

Kelly T. Bocanegra-González https://orcid.org/0000-0001-7177-5856
Thomas S. Prieto-González https://orcid.org/0009-0002-8657-607X
Martin Pullan https://orcid.org/0000-0002-4452-6026
Kyle G. Dexter https://orcid.org/0000-0001-9232-5221
Hannah Atkins https://orcid.org/0000-0002-3523-1883
R. Toby Pennington https://orcid.org/0000-0002-8196-288X

Data availability

All of the data that support the findings of this study are available in the main text or Supplementary Information.

References

Baker TR, Pennington RT, Dexter KG, Fine PVA, Fortune-Hopkins H, Honorio EN, Huamantupa-Chuquimaco I, Klitgård BB, Lewis GP, de Lima HC, Ashton P, Baraloto C, Davies S,

Donoghue MJ, Kaye M, Kress WJ, Lehmann CER, Monteagudo A, Phillips OL, Vasquez R (2017) Maximising synergy among tropical plant systematists, ecologists, and evolutionary biologists. Trends in Ecology & Evolution 32(4): 258–267. https://doi.org/10.1016/j.tree.2017.01.007

Beentje HJ (2010) The Kew plant glossary: an illustrated dictionary of plant terms. Royal Botanic Gardens, Kew, 1–160.

Bisby FA (1970) The evaluation and selection of characters in angiosperm taxonomy: An example from Crotalaria. The New Phytologist 69(4): 1149–1160. https://doi.org/10.1111/j.1499-8137.1970.tb02495.x

Bocanegra-González KT, Dexter KG, Pennington TR, Cuevas-González JP (2024) The genus *Inga* in the Chocó region. University of Tolima Press, Ibagué. https://repository.ut.edu.co/handle/001/3862

Bocanegra-González KT, Dexter KG, Kidner C, Fernández-Méndez F, Giraldo-Gómez S, Camargo-Roa S, de Stapf MS, Pennington RT (2025) Integrative taxonomy recognises four species within the legume *Inga chocoensis* (Leguminosae: Caesalpinioideae). Taxon 74(5): 1091–1105. https://doi.org/10.1002/tax.13375

Brown MT, Wicker LR (2000) Discriminant analysis. In: Tinsley HEA, Brown SD (Eds) Handbook of applied multivariate statistics and mathematical modelling. Academic Press, San Diego, 209–235. https://doi.org/10.1016/B978-012691360-6/50009-4

Butcher B, Smith BJ (2020) Feature engineering and selection: A practical approach for predictive models. The American Statistician 74(3): 308–309. https://doi.org/10.1080/00031305.2020.1790217

Carcassi G, Aidala CA, Barbour J (2021) Variability as a better characterization of Shannon entropy. European Journal of Physics 42(4): 045102. https://doi.org/10.1088/1361-6404/abe361

Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A (2017) Going deeper in the automated identification of herbarium specimens. BMC Evolutionary Biology 17(1): 181. https://doi.org/10.1186/s12862-017-1014-z

Dexter KG, Pennington TD, Cunningham CW (2010) Using DNA to assess errors in tropical tree identifications: How often are ecologists wrong and when does it matter? Ecological Monographs 80(2): 267–286. https://doi.org/10.1890/09-0267.1

Eades DC (1965) The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance. Systematic Zoology 14(2): 98. https://doi.org/10.2307/2411731

Endara MJ, Jaramillo JL (2011) The Influence of Microtopography and Soil Properties on the Distribution of the Speciose Genus of Trees, *Inga* (Fabaceae:Mimosoidea), in Ecuadorian Amazonia. Biotropica 43(2): 157–164. https://doi.org/10.1111/j.1744-7429.2010.00668.x

Eskov VM, Eskov VV, Vochmina YuV, Gorbunov DV, Ilyashenko LK (2017) Shannon entropy in the research on stationary regimes and the evolution of complexity. Moscow University Physics Bulletin 72(3): 309–317. https://doi.org/10.3103/S0027134917030067

GBIF.org (2024) GBIF Occurrence Download. [Downloaded on 31 October 2024] https://doi.org/10.15498/dl.yezzdn

Genero EL, Mario I, Para P, Flora LA (2017) El género *Inga* (Leguminosae: Mimosoideae) del sur de México y Centroamérica, estudio previo para la Flora Mesoamericana. Annals of the Missouri Botanical Garden 80(1): 223–269. https://doi.org/10.2307/2399826

Hagedorn G, Rambold G, Martellos S (2010) Types of identification keys. In: Nimis PL, Vignes Lebbe R (Eds) Tools for identifying biodiversity: Progress and problems. Edizioni Università di Trieste, Trieste, 59–64.

Hall AV (1967) Methods for demonstrating resemblance in taxonomy and ecology. Nature 214(5090): 830–831. https://doi.org/10.1038/214830a0

Hart AG, Bosley H, Hooper C, Perry J, Sellors Moore J, Moore O, Goodenough AE (2023) Assessing the accuracy of free automated plant identification applications. People and Nature 5(3): 929–937. https://doi.org/10.1002/pan3.10490

Hauke J, Kossowski T (2011) Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae 30(2): 87–93. https://doi.org/10.2478/v10117-011-0021-1

Izenman AJ (2008) Modern multivariate statistical techniques. Springer, New York. https://doi.org/10.1007/978-0-387-78189-1

Leon J (1966) Central American and West Indian species of Inga (Leguminosae). Annals of the Missouri Botanical Garden 53(3): 265–359. https://doi.org/10.2307/2394958

Lubischew AA (1962) On the use of discriminant functions in taxonomy. Biometrics 18(4): 455. https://doi.org/10.2307/2527894

Łysko A, Popiela A, Forczmański P, Lukács BA, Barta Z, Maćków W, Wolski GJ (2022) Comparison of discriminant methods and deep learning analysis in plant taxonomy: A case study of Elatine. Scientific Reports 12(1): 20450. https://doi.org/10.1038/s41598-022-24960-1

Mäder P, Boho D, Rzanny M, Seeland M, Wittich HC, Deggelmann A, Wäldchen J (2021) The Flora Incognita app – Interactive plant species identification. Methods in Ecology and Evolution 12(7): 1335–1342. https://doi.org/10.1111/2041-210X.13611

Ohkawa C (2000) Development of teaching materials for field identification of plants and analysis of their effectiveness in science education. The American Biology Teacher 62(2): 113–123. https://doi.org/10.2307/4450850

Palow DT, Nolting K, Kitajima K (2012) Functional trait divergence of juveniles and adults of nine *Inga* species with contrasting soil preference in a tropical rain forest. Functional Ecology 26(5): 1144–1152. https://doi.org/10.1111/j.1365-2435.2012.02019.x

Pennington TD (1997) The genus *Inga*. Royal Botanic Gardens, Kew, Richmond.

Rakouth HN, Andriambololonera S, Rakouth B, Phillipson PB, Lowry PP II, Wilding N (2025) A digital multi-access key for easy identification of large tree species of ebony wood in Madagascar. PhytoKeys 253: 103–132. https://doi.org/10.3897/phytokeys.253.134319

Romero C (2005) Revisión de las especies colombianas de *Inga* sección *Pseudinga*. Estudios en leguminosas colombianas, 174.

Romero C, Alba-López A (2005) Taxonomía del género *Inga* Mill. In: Forero E, Romero C (Eds) Estudios en leguminosas colombianas.

Särkinen T, Hilgenhof R, Gouvêa YF, Giacomin LL, Stone J, Aubriot X, Tepe EJ, Barboza G, Chiarini F, Stern S, Tovar JD, Bohs L, Martine CT, Orejuela A, Orozco CI, Peralta IE, da Silva Sampaio V, Rodríguez A, Knapp S (2025) Taming the beasts: Challenges of identification in big plant genera. Annals of Botany. https://doi.org/10.1093/aob/mcaf164

Schober P, Boer C, Schwarte LA (2018) Correlation coefficients: Appropriate use and interpretation. Anesthesia and Analgesia 126(5): 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Soto J, Pacheco D, Zambrano O, Ortega J (2021) Revisión florística del género *Inga* Miller (Leguminosae-Mimosoideae) en el estado Zulia, Venezuela. Acta Botanica Venezuelica 35(1): 27–52.

Talent N, Dickinson RB, Dickinson TA (2014) Character selection during interactive taxonomic identification: "Best characters.". Biodiversity Informatics 9(1): 1–12. https://doi.org/10.17161/bi.v9i1.4911

Wäldchen J, Rzanny M, Seeland M, Mäder P (2018) Plant species identification using computer vision techniques: A systematic literature review. Archives of Computational Methods in Engineering 25(2): 507–543. https://doi.org/10.1007/s11831-016-9206-z

Supplementary material 1

Morphological characters recorded for *Inga*

Authors: Kelly T. Bocanegra-González, Thomas S. Prieto-González, Martin Pullan, Kyle G. Dexter, Hannah Atkins, R. Toby Pennington

Data type: docx

Explanation note: 74 morphological characters recorded for *Inga*, including their possible states and indication of which were included in the analysis.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/phytokeys.267.172490.suppl1

Supplementary material 2

Categorical characters were standardised following Beentje's Plant Glossary (2010) and Pennington's (1997) taxonomic treatment of *Inga*

Authors: Kelly T. Bocanegra-González, Thomas S. Prieto-González, Martin Pullan, Kyle G. Dexter, Hannah Atkins, R. Toby Pennington

Data type: docx

Explanation note: Illustrated glossary of categorical characters for *Inga*, showing standardised states (e.g., twig shapes, stipule forms, petiole and rachis types, indumentum categories, nectary forms, inflorescence types, calyx shapes, and fruit morphologies).

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/phytokeys.267.172490.suppl2